Available online at www.sciencedirect.com
ScienceDirect
Procedia Computer Science 45 (2015) 3 – 12
Procedia Computer Science

International Conference on Advanced Computing Technologies and Applications (ICACTA-2015)

Predicting User's Web Navigation Behavior Using Hybrid Approach

Meera Narvekar[a] , Shaikh Sakina Banu[b]

[a]*Department of Computer Engineering,D.J.Sanghvi College of Engineering,Mumbai, India.*
[b]*Department of Computer Engineering,D.J.Sanghvi College of Engineering,Mumbai, India.*

Abstract

World Wide Web is growing rapidly with different kinds of websites making it complex along with increasing traffic on the web. However predicting what the user wants becomes very difficult .There are various prediction challenges which are faced , some of them includes long training time, more prediction time, low prediction accuracy, memory limitation etc. The System aims to increase the prediction accuracy particularly when there are many prediction models to consult. The System also aims to reduce the complexity of prediction and yield efficient result and make the prediction user friendly as well minimize Miss-Prediction. The Hybrid model developed combines Markov model as well as Hidden Markov Model which gives user the list of web pages of their interest. We have used various kinds of datasets to analyze, compare and show the effectiveness of Hybrid model using various parameters such as Accuracy, Precision and Miss-Prediction.

Peer-review under responsibility of scientific committee of International Conference on Advanced Computing Technologies and Applications (ICACTA-2015).

Keywords: Prediction models ;Two Tier Architecture ;web mining;N-gram sequence ;Hybrid Model,Dempsters rule,Markov Model,Hidden Markov Model.

1. Introduction

Data mining is a process to extract information from a data set and transform it into better and understandable structure for further use. In other words, Data Mining attempts to extract patterns from the available data in efficient way. Large amount of data has to be interpreted to acquire knowledge about the tasks that occur in the given dataset

Peer-review under responsibility of scientific committee of International Conference on Advanced Computing Technologies and Applications (ICACTA-2015).
doi:10.1016/j.procs.2015.03.073

or database. Web Mining is a technique of data mining whose task is related to the database to collect the relevant information related to the web page access in the form of web session for the future analysis.

In Fig. 1 a general classification of web mining is shown which gives a better view of the structure. Web usage mining helps in better prediction of web page access and the analysis of users' behavior over the World Wide Web. The Internet gives too much of information for users to access and to retrieve various kinds of information as per the needs of the user. It also makes it easy for a user to wander for information as internet contains relevant as well as irrelevant information. Many different patterns in the log record can be used to predict future events and actions for the user. Prediction tries to form various different patterns that help it to predict the next set of actions the user should perform given the available input dataset. Prediction is very efficient to do as more information is gained through search engines and web servers. Prediction can be used at various places over the internet with various different purposes. World Wide Web prediction can also improve searches on many search engines and websites. Another best application of web prediction is recommendation systems, in which we try to find the top n users having the same interests or tastes to a target user record and focus on those set of people to get more traffic on the website and give users best recommendation.

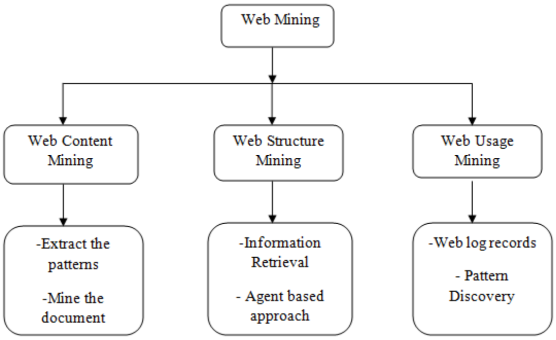

Fig. 1: Taxonomy of web Mining

Web Page Prediction is a problem in which we try to predict the next set of Web pages that a user may visit based on the knowledge of the previously visited web pages. Such kind of knowledge of user's history of browsing web pages within a period of time is referred to as a session[15]. Such log of sessions which are obtained from user's history provide the source of data for training and are extracted from the logs of the Web servers which are stored in a particular format. The log of web server contains sequences of pages that the users have visited in a particular interval of time and date on which it has visited. In Fig. 2 an in general diagram of web page prediction can be seen in which it shows how input is given to prediction model as well as how output is generated which is set of web pages. An array or sequence of web pages is taken as the input such as pages P1, P2, P3, P7 and P6. These set of web pages is given to prediction model which checks for the input and based on the previous knowledge of visiting of these web pages it predicts the next set of web pages which the user can visit. The output of prediction is given as set of web pages which is based on priority and ranking so that the user can visit any web page from the set of web pages as the next web page. In Fig.2, the predicted output is given as page P4 and page P8, based on the analysis of prediction models being used, the user needs to visit pages P4 and P8 after browsing web pages P1, P2, P3, P7 and P6. The accuracy of predicted web pages completely depends on the prediction model being used by the user as well as on good training given to the server logs.

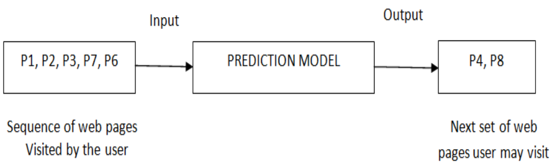

Fig. 2: General Diagram of Web Page Prediction

Prediction models can be classified into two categories such as point based models and path based models[6, 14]. Point-based prediction models are built on actions that are based on time instants and are used to predict the user's next action based on the current actions and behavior. Path-based prediction models are based on user's previous and historic path data or previous history, while point-based prediction is based on currently observed actions or behavior. Accuracy of point-based is low as relatively small amount of information could be extracted from each session to build the prediction model as only current state is considered for prediction .These models draw relatively small amount of information from each session and therefore the prediction can potentially be rather inaccurate. Such kind of applications has a lot of tradeoffs between complexity of prediction and prediction accuracy. Thus a system is developed which can do prediction accurately and give result in less time so that the system proves to be efficient than many existing system that is used for prediction. We aim to reduce the prediction time particularly when there are many prediction models to consult. We also aim to reduce the complexity of prediction and yield efficient result and make the prediction user friendly. The system which is developed will give better accuracy than the existing systems.

2. Literature Review

Web Page Prediction is used to predict next set of web pages that are required by the user .Prediction of user's behavior is based on user's previous history and knowledge. Internet is so large and complex that user gets confused in the web and is unable to get proper result. However prediction has made it simple for user to get the required result more efficiently from the web. There are various techniques developed for better prediction of users' behavior and to give user the next set of web pages that the user can use for navigation. Priyanka Makkar, Payal Gulati, Dr. A.K. Sharma[1] proposed a novel approach for increasing web performance by analyzing and predicting user behavior both by collaborating information from user access log and website structure repository. However, there are some drawbacks of using this technique such as by getting the whole path for website structure, it consumes memory and makes the proposed system degrade performance. Mamoun A. Awad and Latifur R. Khan[2] proposed a system which uses several hybrid models that combines different techniques such as Markov model, artificial neural networks (ANNs), and the All-Kth-Markov model, which is used for prediction using Dempster's rule. However, Markov Model cannot predict a session that does occur in the training set, because such session will have zero probability as it is not there in dataset and we cannot find probability.

Bamshad Mobasher, Honghua Dai, Tao Luo, Miki Nakagawa [3]proposed the frequent item set graph to match an active user session with frequent item sets and predict the next page that the user is likely to visitHowever, the computation time taken by the framework is very large and it makes the system complex. Xing Dongshan and Shen Junyi[4] proposed a new markov model which can minimize user-perceived latency, which is crucial in the rapidly growing World Wide Web. However, it sometimes lacks in accuracy of the result.Zhong Su, Qiang Yang, Ye Lu, Hong jiang Zhang[5] a system whose result can potentially be applied to a wide range of applications on the web, including pre-sending, pre-fetching, enhancement of recommendation systems as well as web caching policies. They compared the effectiveness of n-gram prediction for different sequence length n, and found that with an increase in sequence length, there is an increase in precision and decrease in applicability.

Sawan Bhawsar, Kshitij Pathak, Vibhor Patidar[6] proposed a system in which they have used page rank algorithm and matrix to predict the users' behavior. Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava[7] proposed a technique called as webminer, which is used for better prediction of web browsing behavior. Also, a method to divide user sessions into semantically meaningful transactions is defined and successfully tested against two other methods[8-14]. Mamoun A. Awad and Issa Khalil[15] proposed a new modified Markov model to alleviate the issue of scalability in the number of paths.

3. Architectural Diagram

In Fig. 3, the user will enter the query or choice list as per his interest. The query which the user enters will go to the prediction model which will analyze the request and will give the result as per user query as well as update the database for better prediction. The request of the user will go to the database where the search will take place and the response along with other choices will be returned to the user.

The functionality of various modules are given below:

- User: The person who needs optimized output from the proposed system. It may be customer, stakeholder or a tester of the system.
- Choice: The query which the user wants to search. The choice will always be optimized choice according to the user's request.
- Choice List: Choice list is returned to the user as per previous knowledge of his/her visit to different web pages. Choice list is generated by the prediction model using previous logs and dynamic database.

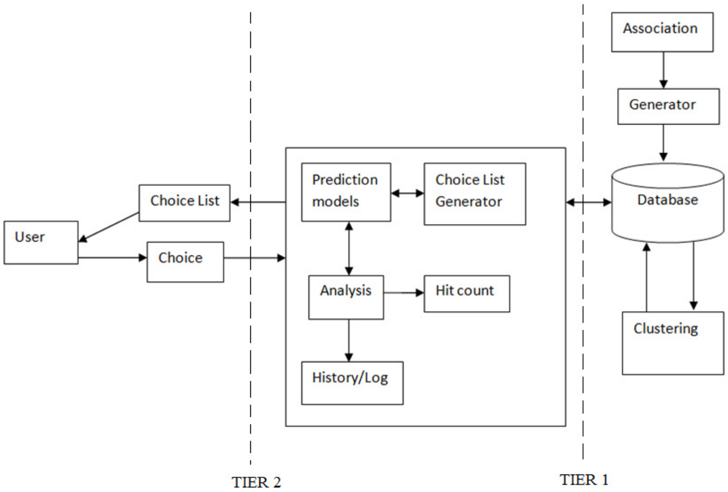

Fig. 3: Architectural Diagram

- Prediction Models: Prediction models are used for better prediction of next web page the user wants to visit. Prediction model helps to give user better choices in less time so reduces the complexity of the system. Prediction models are used for better prediction of next web page the user wants to visit. Prediction model helps to give user better choices in less time so reduces the complexity of the system. The various prediction models used in our system are Markov Model, Hidden Markov Model. Prediction models first checks from the memory if the next set of web pages are available, if not available then it considers the training provided by the database. We have used Markov Model, Hidden Markov model and combination of both as well.
- Choice List Generator: It generates the list of choices from the dynamic database. This choice list generator is compared with previous logs of the user to give the user an optimized result. The list is presented from highest priority to the lowest priority.
- Database: The database is generated dynamically using associations and clustering. The database will have processed data which will be in the form as per required by the system to process it.

3.1. System Flow Diagram

Our systems consist of two tier framework wherein tier 1 is done offline where training of database is done using backward and forward probability. While tier 2 is done online based on training. Prediction is done at tier 2.Two Tier architecture is being used so that any changes to be done at any tier can be done without disturbing any other tier and it also reduces the complexity of the system. Whenever user enters a choice for searching particular information, that choice is given as test data to the tier 2 of the system. Tiers 2 consist of prediction module which will give the output of the choice given by the user. Tier 2 will get the trained data from the tier 1 which stores all the sessions in a specific format. Tier 2 will use the information given by the user as well as the training dataset and give the user best result of prediction.

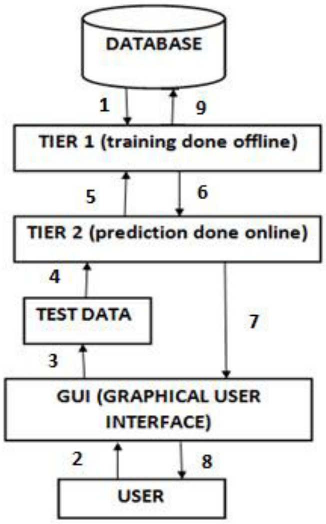

Fig. 4 : System Flow

3.2. Prediction models

3.2.1. Markov Model

The basic concept of Markov Model is to predict the next set of web pages depending on the result of previous history of the user. In Web prediction, the next action or behavior corresponds to predicting the next set of web page to be visited by the user based on his previous sessions. The previous action of the user corresponds to the previous set of web pages that have already been browsed. In Web page prediction, the K^{th}-order Markov model is given as the probability that a user will visit the K+1th page provided that the user has visited the ordered K^{th} web pages[15]. For example, in the second-order Markov model, prediction of the next set of web page is computed based only on the previous two web pages previously visited or browsed. Markov Model repeatedly searches for direct relation between web pages.

Consider S1, S2, S3...Sn be the set of web pages visited by the user and s is the sequence of pages we need to predict $Sn+1^{th}$ page, so the probability is given as:

Pr (Sn+1=s | S1=s1, S2=s2…….Sn=sn) (1)

If first order Markov model is used for prediction then the probability is given as:

Pr (S2=s | S1=s1). (2)

For second order Markov model, the probability is given as:

Pr (S3=s | S2=s2, S1=s1) (3)

Similarly all other orders of Markov models are calculated.

The main advantages of Markov model is its efficiency and performance in terms of building a model. It can be easily shown that building various orders of Markov model is linear with the size of the training set provided. The key idea is to use good data structure to build and keep track of each pattern along with its probability, for example hash tables, heap etc. Prediction is performed in consistent time because the running time of accessing an entry in a hash table is constant[4]. If the numbers of states are fixed and there is no need to change the domain for searching the web page then Markov model can be used for prediction. Markov Model does not give better results in case of cross domain searching of information.

3.2.2. Hidden Markov Model

Hidden Markov model is a statistical model and an intelligent system where it considers even the state sequence through which the model passes. There is no one-to-one correspondence between the states. The states can keep on changing as it considers hidden states as well[16]. Consider a sequence of web pages as Y= y1, y2, y3···. yn-1. The probability of next sequence is given as Pr (Y) = Σ (for all x) P (Y|X) P(X) , where the hidden sequence of web pages is given as X= x1,x2,x3······xn-1.

Transition Matrix is used to get relation or association between states and the content in the states, whereas Emission Matrix is used to get relation or association between various states. Emission and Transition Matrix is developed by the Tier 1 where training of dataset is done. Hidden Markov Model adapts to changes in the behavior of the system. Since the database is very large and it is difficult to handle large database, Hidden Markov Model works well with large database. It learns the behavior of the user well before predicting so that the result of prediction is efficient.

3.2.3. Dempsters Rule

Dempster's rule is a well known method for aggregating many different bodies of evidence in the same reference set as the input given to the dempsters rule should not be in different format[6]. Suppose we want to combine various evidences for a hypothesis C. In Web page prediction, C is the current page during prediction for a user sessions. For example, we want to know what is the next page a user might visit after visiting pages p1, p3, p4, and p9. C is a member of 2 Θ, which is the power set of Θ, where Θ is said to be the frame of discernment. A frame of discernment is a set of mutually exclusive elements which is exhaustive. Given two independent set of sources of evidences, m11 and m12, Dempster's rule combines them in the following structure of frame [6, 14]:

$$M_{11,12}\,(C)=M_{(A,B\subseteq\theta,A\cap B=C)}\,M11(A)M12(B)/M_{(A,B\subseteq\theta,A\cap B\neq\emptyset)}\,M11(A)M12(B) \qquad (4)$$

Here, A and B are supersets of C. M11 and M12 are functions (also known as a mass of belief) that assign a coefficient between 0 and 1.M11,12(C) is the combined Dempsters–Shafer probability for a hypothesis C.

3.2.4. Hybrid Model

By observing the behavior of Markov and Hidden Markov Model, we have seen that Markov Model works better for 1gram sequence only and sometimes for 2 gram sequence, however if we increase the number of N-grams ,the prediction accuracy decreases drastically which sometimes lead to no output. While Hidden Markov Model works best for huge database and it always gives an optimized output for higher values of N-grams. We combine the results of Markov as well as Hidden Markov Model into Dempster's rule[17], where it takes the probability and based on decision parameter gives the prediction list as the output which is as per the requirement of the user and its behavior.

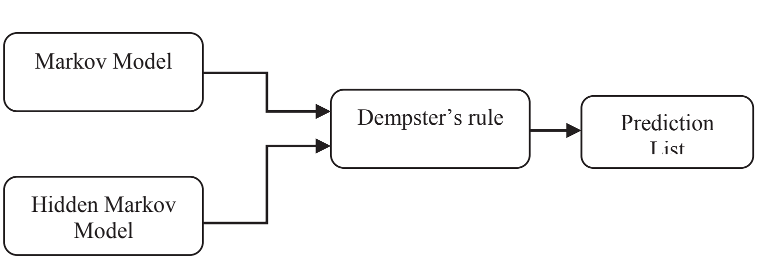

Fig 5: Hybrid Model

4. Result and Analysis

We have used Markov model, Hidden Markov model and combination of both in hybrid system for prediction. To measure better accuracy we have done training and testing of the processed dataset .We have considered 65% for training and 35% for testing purpose to get better prediction. We have tested the dataset many number of times to check for accuracy of developed system .For each training and testing purpose we have considered N grams .For example, first we have trained and tested dataset for 1 gram and then for 2 grams and so on. Maximum we have considered 6 grams for representation purpose. The more the number of N-grams, the more the accuracy keeps on decreasing. We keep on training the dataset for each number of N- grams on several numbers of iterations and check for correctness of the result. We have considered two parameters for showing that the result is better than the existing system that is Accuracy and Miss Prediction.

Table 1: Correct prediction of Markov Model, Hidden Markov Model (HMM) and Hybrid Model (HYB) for NASA July ’1995 Dataset

NASA Dataset July 1995											
Dataset 1				**Dataset 2**				**Dataset 3**			
Markov	**HMM**	**HYB**	**Total**	**Markov**	**HMM**	**HYB**	**Total**	**Markov**	**HMM**	**HYB**	**Total**
100	90	104	350	60	50	60	400	130	120	130	310
80	90	94	310	60	50	64	350	120	180	184	400
100	120	124	340	40	50	52	290	150	190	196	600
100	110	114	350	80	100	102	400	160	170	174	600
90	100	104	320	100	90	90	400	170	190	196	550
80	90	94	300	80	90	98	450	60	70	70	310

4.1. Accuracy

Accuracy is defined as total number of correct predictions per total set of records.
Accuracy = number of correct prediction /total number of records.

Table 2: Accuracy table for Markov Model, Hidden Markov Model and Hybrid Model of July 1995 dataset.

	Accuracy(July 1995 Dataset)								
	Dataset 1			**Dataset 2**			**Dataset 3**		
	Markov	**HMM**	**HYB**	**Markov**	**HMM**	**HYB**	**Markov**	**HMM**	**HYB**
1-Gram	0.2857	0.2571	0.2971	0.1765	0.125	0.15	0.4194	0.3871	0.4194
2-Gram	0.2581	0.2903	0.3032	0.1714	0.1429	0.1829	0.3	0.45	0.46
3-Gram	0.2941	0.3529	0.3647	0.1379	0.1724	0.1793	0.25	0.3167	0.3267
4-Gram	0.2857	0.3143	0.3257	0.2	0.25	0.255	0.2667	0.2833	0.29
5-Gram	0.2813	0.3125	0.325	0.25	0.225	0.225	0.3091	0.3455	0.3564
6-Gram	0.2667	0.3	0.3133	0.1778	0.2	0.2178	0.1935	0.2258	0.2258

The graph of accuracy for dataset 1, dataset 2 and dataset 3 is given as:

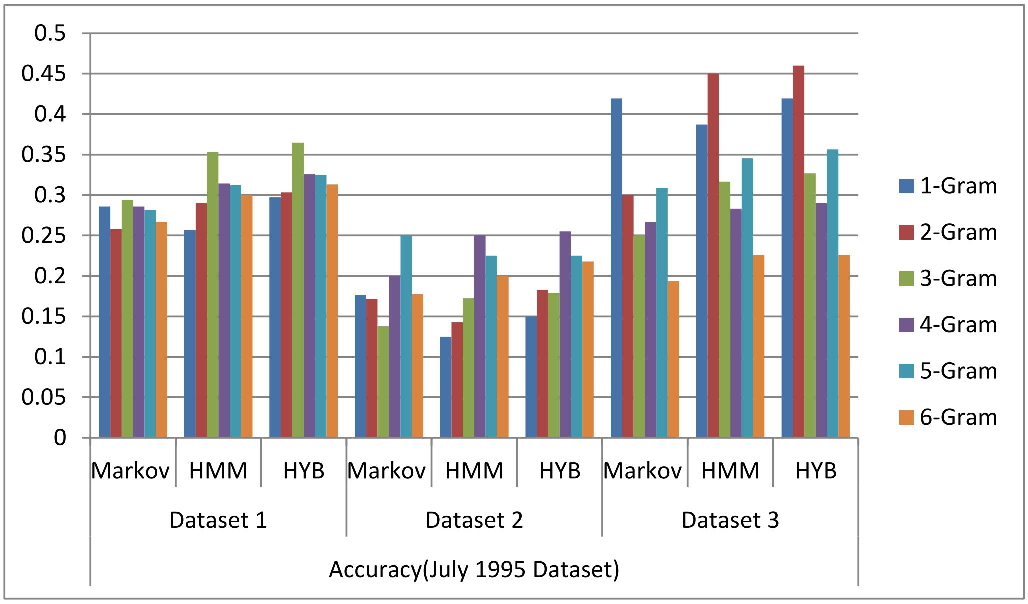

Fig. 6: Accuracy for July 1995 NASA datasets using Markov model, Hidden Markov model and Hybrid model

Accuracy is based on the proper prediction of web pages of a particular user. However the behavior of user keeps on changing at every instinct. Since we are working with loosely coupled data, there will be little or no relations between the web pages .The model is developed in such a way that user will always get better and accurate prediction on loosely coupled data. In fig. 6, the accuracy value keeps on changing depending on the relations between various web pages given in the datasets of NASA.

4.2. Miss Prediction

A predicted output cannot be considered as incorrect as the algorithm or the prediction model would give some output which will be related in some or the other way, thus we cannot say that prediction is incorrect. Such a scenario can be said as miss prediction. Let Pr be the correctly detected values of prediction.

Miss Prediction is given as: Miss Prediction= 1- Pr (Correct Prediction i.e. Precision).

Precision is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved.

Table3: Precision calculated for July 1995 NASA Dataset for Markov, Hidden Markov Model and Hybrid Model.

	Precision (July 1995)								
	Dataset 1			***Dataset 2***			***Dataset 3***		
	Markov	***HMM***	***HYB***	***Markov***	***HMM***	***HYB***	***Markov***	***HMM***	***HYB***
1-Gram	0.4	0.3462	0.4228	0.1765	0.1429	0.1765	0.7222	0.6316	0.7222
2-Gram	0.3478	0.4091	0.4352	0.2069	0.1667	0.2238	0.4286	0.8182	0.8519
3-Gram	0.4167	0.5455	0.5741	0.16	0.2083	0.2185	0.3333	0.4634	0.4851
4-Gram	0.4	0.4583	0.4831	0.25	0.3333	0.3423	0.3636	0.3953	0.4085
5-Gram	0.3913	0.4545	0.4815	0.3333	0.2903	0.2903	0.4474	0.5278	0.5537
6-Gram	0.3636	0.4286	0.4563	0.2162	0.25	0.2784	0.24	0.2917	0.2917

Table 4: Miss Prediction calculated for July 1995 NASA Dataset using Markov, Hidden Markov Model and Hybrid model.

	Miss Prediction (July 1995)								
	Dataset 1			Dataset 2			Dataset 3		
	Markov	HMM	HYB	Markov	HMM	HYB	Markov	HMM	HYB
1-Gram	0.6	0.6538	0.5772	0.8235	0.8571	0.8235	0.2778	0.3684	0.2778
2-Gram	0.6522	0.5909	0.5648	0.7931	0.8333	0.7762	0.5714	0.1818	0.1481
3-Gram	0.5833	0.4545	0.4259	0.84	0.7917	0.7815	0.6667	0.5366	0.5149
4-Gram	0.6	0.5417	0.5169	0.75	0.6667	0.6577	0.6364	0.6047	0.5915
5-Gram	0.6087	0.5455	0.5185	0.6667	0.7097	0.7097	0.5526	0.4722	0.4463
6-Gram	0.6364	0.5714	0.5437	0.7838	0.75	0.7216	0.76	0.7083	0.7083

The graph obtained shows the comparison of various datasets using markov model, hidden markov model and hybrid model and it can be seen that hybrid model gives better result than markov model and hidden markov model as hybrid system combines the two models and generates the output. As we see in table 3 and table 4, by using the Hybrid model Accuracy value keeps on increasing as we increase the N-gram sequence, whereas Miss Prediction keeps on decreasing with the increase in N-gram sequence.

The graph for Miss Prediction on Dataset1, Dataset2 and Dataset3 is given as below:-

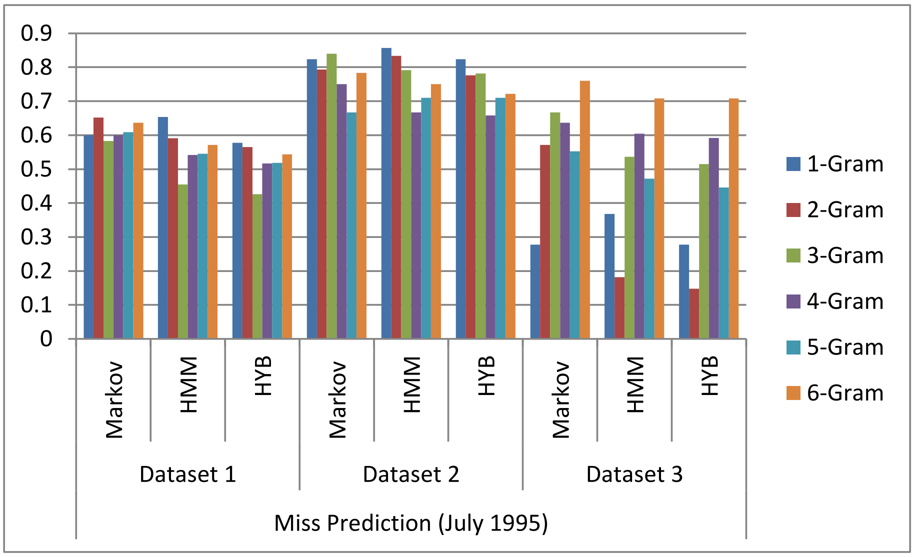

Fig 7. Miss Prediction graph calculated for July 1995 NASA datasets using Markov model, Hidden Markov model and Hybrid model

5. Conclusion

Web Mining is the task related to the database to collect the information related to the web page access in the form of web session for the future analysis. Prediction is a problem in which we need to predict next set of web pages which the user will visit knowing all the previous visits of the user in a session. However exact prediction model to predict accurately is not built till yet, so the proposed system will overcome many of the drawbacks taking into consideration that the performance will not degrade.

The system improves prediction accuracy without compromising prediction time. The system consist of two Tiers wherein Tier one is used for training while Tier two is used for prediction. We performed experiments on NASA dataset using Markov model, Hidden Markov model as well as Hybrid model with parameters such as accuracy, miss-prediction and maximum number of N-grams. The hybrid model works better for higher N-gram sequence and gives better accuracy and less miss prediction. The model studies the human behavior and gives prediction as per the user requirement, thus increasing the traffic on one's website as well as helping the server to manage the resources efficiently. In the future, we plan to convert the two tier architecture to three Tier Architecture by introducing a classifier in between existing two tier architecture which will decrease prediction time. Classifier should be selected in such a way that it should give quick response about the type of model to be selected so that time is consumed less as well as always the best model is selected for prediction.

Acknowledgements

All the authors would like to thank head, department of computer engineering (DJSCOE). I would also like to thank research guide Meera Narvekar for her valuable comments and help towards this research work.

References

1.Priyanka Makkar, Payal Gulati, Dr. A.K. Sharma, " A Novel Approach for Predicting User Behavior for Improving Web Performance ".International Journal on Computer Science and Engineering ,Vol. 02, No. 04, 2010, 1233-1236.
2. M. Awad and L. Khan, "Web navigation prediction using multiple evidence combination and domain knowledge," IEEE Trans. Syst., Man,Cybern. A, Syst., Humans, vol. 37, no. 6, pp. 1054–1062, Nov. 2007.
3. B. Mobasher, H. Dai, T. Luo, and M. Nakagawa, "Effective personalization based on association rule discovery from Web usage data," in Proc. ACM Workshop WIDM, Atlanta, GA, Nov. 2001.
4. Xing Dongshan And Shen Junyi , Xian Jiaotong University," A New Markov Model For Web Access Prediction".
5. Z. Su, Q. Yang, Y. Lu, and H. Zhang, "WhatNext: A prediction system for Web requests using n-gram sequence models," in Proc. 1st Int. Conf. Web Inf. Syst. Eng. Conf., Hong Kong, Jun. 2000, pp. 200–207.
6. M. Awad, L. Khan, and B. Thuraisingham, "Predicting WWW surfing using multiple evidence combination," VLDB J., vol. 17, no. 3, pp. 401– 417, May 2008.
7. Sawan Bhawsar, Kshitij Pathak, Vibhor Patidar ,"New Framework for Web Access Prediction", International Journal of Computer Technology and Electronics Engineering (IJCTEE) ,Volume 2 , Issue 1.
8. R. Cooley, B. Mobasher, and J. Srivastava, "Data preparation for mining World Wide Web browsing patterns," J. Knowl. Inf. Syst., vol. 1, no. 1, pp. 5–32, 1999.
9. Sampath P.and Ramya D, Performance Analysis Of Web Page Prediction With Markov Model, Association Rule Mining(Arm) And Association Rule Mining With Statistical Features(Arm-Sf). IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727Volume 8, Issue 5 (Jan. - Feb. 2013), PP 70-74.
10. R. Agrawal and R. Srikant, "Fast algorithms for mining association rules,"in Proc. 20th Int. Conf. VLDB, Santiago, Chile, 1994.
11. B. M. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Analysis of recommender algorithms for e-commerce," in Proc. 2nd ACM EC,Minneapolis, MN, Oct. 2000.
12. O. Nasraoui and C. Petenes, "Combining Web usage mining and fuzzy inference for Website personalization," in Proc. WebKDD, 2003, pp. 37–46.
13. M. T. Hassan, K. N. Junejo, and A. Karim, "Learning and predicting key Web navigation patterns using Bayesian models," in Proc. Int. Conf. Comput. Sci. Appl. II, Seoul, Korea, 2009, pp. 877–887.
14. I. Zukerman, W. Albrecht, and A. Nicholson, "Predicting user's request on the WWW," in Proc. 17th Int. Conf. UM, 1999, p. 393.
15. Mamoun A. Awad and Issa Khalil ,"Prediction of User's Web-Browsing Behavior: Application of Markov Model", IEEE transactions on systems, man, and cybernetics—part b: cybernetics, vol. 42, no. 4, august 2012.
16. Rabiner, L. Juang, B.H. "An introduction to hidden Markov models" ISSN :0740-7467, IEEE Xplore, 1986.
17. G.Shafer, "A mathematical theory of Evidence" , Princeton Univ , press ,1999.